Naïve Bayes classifier

A small weather data set on previous records of

(i) weather conditions and

(ii) **whether certain event happens** (i.e. certain activity is "played")

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

A new case for prediction:

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

Play=?

- **E**: The evidences (observations) we have:

  $E_1 =$ Outlook = Sunny, $E_2 =$ Temp. = Cool, $E_3 =$ Humidity = High, $E_4 =$ Windy = True

  E = (E_1, E_2, E_3, E_4) =

Outlook	Temp.	Humidity	Windy
Sunny	Cool	High	True

- **H**: whether the event happens:
  Two possible predictions (hypotheses):

  H = "Play=yes" or

  H = "Play=no"

The basic probabilistic approach:

Compare two conditional probabilities

Pr(H = “Play=yes” | E = | Outlook | Temp. | Humidity | Windy |
|---|---|---|---|
| Sunny | Cool | High | True |

) versus

Pr(H = “Play=no” | E = | Outlook | Temp. | Humidity | Windy |
|---|---|---|---|
| Sunny | Cool | High | True |

) .

It is hard to directly estimate these two probabilities from the data set. (Why?)

The Naïve Bayes approach:

(I) Use the Bayes rule

- Probability of event H given evidence E:

$$Pr[H|E] = \frac{Pr[E|H]Pr[H]}{Pr[E]}$$

- *A priori* probability of H : $Pr[H]$
  - Probability of event *before* evidence is seen
- *A posteriori* probability of H : $Pr[H|E]$
  - Probability of event *after* evidence is seen

Thomas Bayes
Born: 1702 in London, England
Died: 1761 in Tunbridge Wells, Kent, England

The Naïve Bayes approach:

Just calculate $Pr[E|H]Pr[H]$ as the likelihood and compare

Pr(E = | Outlook | Temp. | Humidity | Windy |
|---|---|---|---|
| Sunny | Cool | High | True |

| H = “Play=yes”) *

Pr(H = “Play=yes”)

with

Pr(E = | Outlook | Temp. | Humidity | Windy |
|---|---|---|---|
| Sunny | Cool | High | True |

| H = “Play=no”) *

Pr(H = “Play=no”).

No need to worry about $Pr[E]$ since

Pr(E = | Outlook | Temp. | Humidity | Windy |
|---|---|---|---|
| Sunny | Cool | High | True |

) is the same denominator on the right hand side of the Bayes rule.

The Naïve Bayes approach:

(II) Use the Naïve assumption on independency: Individual evidences (E_1, E_2, E_3, E_4 …) are independently affected by the underlying event separately.

$$Pr[E|H]$$

$$= Pr[E_1|H] * Pr[E_2|H] * \ldots * Pr[E_n|H]$$

It is much easier estimate conditional probabilities

$Pr(E_1 \mid H)$, $Pr(E_2 \mid H)$, $Pr(E_3 \mid H)$, and $Pr(E_4 \mid H)$ from the data set . (Why?)

How to estimate $Pr(E_1 | H)$, $Pr(E_2 | H)$, $Pr(E_3 | H)$, $Pr(E_4 | H)$?

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/	5/
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5	14	14
Rainy	3/9	2/5	Cool	3/9	1/5								

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Estimate $Pr(E_1 \mid H)$ and $Pr(E_2 \mid H)$

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/	5/
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5	14	14
Rainy	3/9	2/5	Cool	3/9	1/5								

$Pr(E_1 =$ Outlook = Sunny $\mid H =$ "Play=yes") : 2/9 (why?)

$Pr(E_1 =$ Outlook = Sunny $\mid H =$ "Play=no") : 3/5 (why?)

$Pr(E_2 =$ Temp. = Cool $\mid H =$ "Play=yes") : 3/9 (why?)

$Pr(E_2 =$ Temp. = Cool $\mid H =$ "Play=no") : 1/5 (why?)

Estimate $Pr(E_3 \mid H)$ and $Pr(E_4 \mid H)$

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/	5/
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5	14	14
Rainy	3/9	2/5	Cool	3/9	1/5								

$Pr(E_3 =$ Humidity / High $\mid H =$ "Play=yes") : 3/9 (why?)

$Pr(E_3 =$ Humidity / High $\mid H =$ "Play=no") : 4/5 (why?)

$Pr(E_4 =$ Windy / True $\mid H =$ "Play=yes") : 3/9 (why?)

$Pr(E_4 =$ Windy / True $\mid H =$ "Play=no") : 3/5 (why?)

Estimate Pr(H):

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/	5/
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5	14	14
Rainy	3/9	2/5	Cool	3/9	1/5								

Pr(H= “Play=yes”) = 9/14 (why?)

Pr(H= “Play=no”) = 5/14 (why?)

Just calculate the likelihood $Pr[E|H]Pr[H]$

for H = “Play=yes” and for H = “Play=no”

For example,

Pr(E = | Outlook | Temp. | Humidity | Windy |
|---|---|---|---|
| Sunny | Cool | High | True |

| H = “Play=yes”) *

Pr(H = “Play=yes”)

=

$$Pr[Outlook=Sunny|yes] \times Pr[Temperature=Cool|yes] \times Pr[Humidity=High|yes] \times Pr[Windy=True|yes] \times \frac{Pr[yes]}{}$$

The results:

Likelihood of the two classes

For "yes" = $2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14 = 0.0053$

For "no" = $3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0.0206$

Conversion into a probability by normalization:

P("yes") = 0.0053 / (0.0053 + 0.0206) = 0.205

P("no") = 0.0206 / (0.0053 + 0.0206) = 0.795

The prediction: **"**Play=no"